to situations in which the model allows only for the effect of a single nonnull haplotype, effectively comparing this "target" haplotype with all others. Tests based on such a procedure can perform very poorly in cases where more than one haplotype affects disease risk. We feel that the modeling approach is especially important in just this case, because hypothesis tests for single haplotype effects are tests of a composite null hypothesis; such tests require the capacity to *estimate* relative risk parameters for those haplotypes not constrained by the null hypothesis. But estimation requires modeling the effects of multiple haplotypes, which appears to go beyond the identifiability results presented by Lin and Zeng. When interest is limited to models of a single haplotype, *tests* can be constructed without much difficulty that are valid regardless of the distribution of haplotypes (see, e.g., Schaid et al. 2002; Zaykin et al. 2002). Thus the identifiability results of Lin and Zeng are most important in situations where one wishes to *estimate* the effect of a single haplotype (relative to all of the others). Can these results actually be used to analyze real data? The answer to this question is less clear, because the haplotype distribution parameters may be only weakly identified in finite samples, especially when the true parameters are close to the null hypothesis or where the true risk model is close to dominant (a fact that will not always be known a priori). Along these lines, we note that in their analyses of the FUSION and simulated data, Lin and Zeng use the stronger assumption that model 3 is correct.

Finally, some quibbles. Lin and Zeng claim to describe analytical methods for "all commonly used study designs." In fact, genetic epidemiologists often use family-based association studies, such as case-parent trio studies, that are not covered by Lin and Zeng's article. Recently, we have developed methods for fitting haplotype risk models using case-parent trio data that are robust to misspecification of the parental haplotype distribution (Allen, Satten, and Tsiatis 2005). We have extended our approach to include haplotype–covariate interactions, where the robustness to misspecification of the parental haplotype distribution enables a general dependence of haplotype frequencies on covariates. These methods are based on the efficient score function; we are currently studying the application of our approach to case-control studies. In particular, it appears possible to remove any dependence of the distribution of H given G in models with no covariates; given the assumption that Lin and Zeng were forced to make, this will be of particular interest should these methods extend to models that include haplotype–covariate interactions in case-control studies. Another design also not considered by Lin and Zeng corresponds to conditional logistic regression of finely stratified data. Here we note that the retrospective approach of Epstein and Satten (2003) can be used with highly stratified data because the intercept parameter is conditioned out; as a result, we can use this approach when we have a large number of intercept parameters. We have also developed an extension of the Epstein and Satten approach that includes covariate effects in addition to haplotype effects (and their interactions) for matched or highly stratified studies.

In summary, we congratulate Lin and Zeng on an interesting and stimulating article.

ADDITIONAL REFERENCES

Allen, A. S., Satten, G. A., and Tsiatis, A. A. (2005), "Locally-Efficient Robust Estimation of Haplotype–Disease Association in Family-Based Studies," *Biometrika*, 92, 559–571.

Carroll, R. J., Wang, S., and Wang, C. Y. (1995), "Prospective Analysis of Logistic Case-Control Studies," *Journal of the American Statistical Association*, 90, 157–169.

Comment

Nilanjan CHATTERJEE, Christine SPINKA, Jinbo CHEN, and Raymond J. CARROLL

Lin and Zeng are to be congratulated on an article that describes identifiability and estimation of haplotype distributions and risk parameters for very general models, both prospectively and for case-control studies. In particular, the identifiability conditions will give important guidance to researchers as they attempt to use different models for haplotypes besides Hardy–Weinberg equilibrium (HWE).

Our major aim in this comment is to place Lin and Zeng's article in the broader context of various alternative methods for haplotype-based regression analysis. We point out the connections and the differences between these alternative methods, to shed light on their relative merits. In particular, we note that in some important subproblems, other methods are available. These methods are efficient and simple to implement, and they avoid the need to estimate possibly high-dimensional nuisance parameters.

1. CASE–CONTROL STUDIES

Because haplotype-based association studies are becoming increasingly popular, a number of researchers have developed methods for logistic regression analysis of case-control studies in the presence of phase ambiguity. The methods can be broadly classified into two categories: prospective and retrospective. Before going into technical details, it is useful to understand the main principles behind these two classes of methods.

Nilanjan Chatterjee is Senior Investigator, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20852 (E-mail: *chattern@mail.nih.gov*). Christine Spinka is Assistant Professor, Department of Statistics, University of Missouri, Columbia, MO 65211-6100 (E-mail: *spinkac@missouri.edu*). Jinbo Chen is Research Fellow in the Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20852 (E-mail: *chenjin@mail.nih.gov*). Raymond J. Carroll is Distinguished Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *carroll@stat.tamu.edu*). Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health through a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

In the Public Domain
Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000835

With a slightly different notation than that of Lin and Zeng, let D be disease status, H^d be the "diplotypes" (i.e., the two haplotypes an individual carries in his or her pair of homologous chromosomes), G be the observed genotype, and X be the nongenetic/environmental covariates. Let the risk function satisfy

$$\text{logit}\{\text{pr}(D = 1|H^d, X)\} = \beta_0 + m(H^d, X, \beta_1), \qquad (1)$$

where $m(\cdot)$ is known but of completely general form.

Under the foregoing notation, the prospective likelihood of the data is given by $\text{pr}(D|G, X)$, which ignores the fact that under the case-control sampling design, data are observed on (G, X) conditional on D. In contrast, the retrospective likelihood of the data is given by $\Pr(G, X|D)$ and accounts for the underlying case-control sampling design.

When there are no missing data (i.e., $G = H^d$), it follows from the well-known results of Prentice and Pyke (1979) that the prospective approach is actually equivalent to the retrospective maximum likelihood analysis, provided that the distribution of the covariates (G, X) is treated completely nonparametrically. Thus the prospective method is a "robust approach" for analysis of case-control studies that does not rely on any assumption about the covariate distribution. In studies of genetic epidemiology, however, it often may be reasonable to assume certain parametric or semiparametric models for the covariate distribution in the underlying source population. The assumptions of HWE and gene–environment independence are examples of such models. The retrospective likelihood can directly incorporate such assumptions into the analysis and can be much more efficient than the prospective method when the assumptions are valid (Epstein and Satten 2003; Chatterjee and Carroll 2005).

1.1 Retrospective Maximum Likelihood Analysis With Haplotype-Phase Ambiguity

Epstein and Satten (2003) first described the retrospective maximum likelihood method for haplotype-based association analysis of case-control studies. Incorporation of nongenetic covariates X in this method is complicated by the fact that the retrospective likelihood involves potentially high-dimensional nuisance parameters that specify the distribution of X in the underlying population. In the gene–environment interaction context, and as in the simulation study and example of Lin and Zeng, it is often reasonable to assume that H^d and environmental factors X are independent in the population, with a parametric form

$$\text{pr}(H^d = h^d|X) = \text{pr}(H = h^d) = q(h^d, \theta), \qquad (2)$$

where the model $q(h^d; \theta)$ in turn could be specified according to HWE or some of its extensions, as considered by Lin and Zeng. More generally, one can assume a parametric model for the diplotype distribution of the form

$$\text{pr}(H = h^d|X = x) = q(h^d, x, \theta). \qquad (3)$$

Model (3), for example, can incorporate departure from gene–environment independence and HWE that may be caused by "population stratification." In particular, one could assume HWE and gene–environment independence conditional on various demographic factors, such as ethnicity and geographic regions, and specify the haplotype frequencies conditional on these factors according to a parametric model, such as the polytomous logistic regression model (Spinka, Carroll, and Chatterjee 2005). Moreover, (3) potentially can be used to directly model the association between haplotypes and environmental exposure X.

Under models (2) and (3), Spinka et al. (2005) described simple and easily computable methods that avoid estimating the nonparametric marginal distribution of X and exploit the information available in (2) or (3) to increase efficiency. Chatterjee, Kalaylioglu, and Carroll (2005) described similarly simple methods applicable for family-based or other types of individually matched case-control studies. Let there be n_1 cases and n_0 controls, and let $\pi = \text{pr}(D = 1)$ be the marginal probability of the disease in the population. Assume the definitions

$$\kappa = \beta_0 + \log(n_1/n_0) - \log\{\pi/(1 - \pi)\}$$

and

$$S(d, h, x, \Omega) = q(h, x, \theta)\frac{\exp[d\{\kappa + m(h, x, \beta_1)\}]}{1 + \exp\{\beta_0 + m(h, x, \beta_1)\}},$$

where $\Omega = (\beta_0, \kappa, \theta^{\text{T}}, \beta_1^{\text{T}})^{\text{T}}$. Let $\mathcal{H}_G$ be the set of diplotypes consistent with the observed genotype G. Define

$$L^*(D, G, X, \Omega) = \frac{\sum_{h\in\mathcal{H}_G} S(D, h, X, \Omega)}{\sum_h \sum_d S(d, h, X, \Omega)}.$$

Spinka et al. (2005) first showed that under certain conditions, which are easily verifiable from the data, all of the parameters in Ω, including the intercept parameter β_0, are identifiable from the retrospective likelihood $\prod_i \Pr(G_i, X_i|D_i)$, as long as the underlying models are specified in such a way that Ω would be identifiable from prospective studies. Moreover, the maximum retrospective likelihood estimate of Ω can be obtained as a solution of the score equation corresponding to the pseudolikelihood

$$l^* = \sum_{i=1}^{N} \log\{L^*(D_i, G_i, X_i, \Omega)\}. \qquad (4)$$

Spinka et al. described strategies for estimating the regression parameter β_1 based on l^* for both known and unknown values of the marginal probability of the disease in the underlying population. If one is also willing to make the rare disease assumption *for all* H^d and X, then ℓ^* effectively becomes equivalent to the method that Lin and Zeng derived in their section A.4.5 under the assumption of gene–environment independence. Note, however, that neither the rare disease approximation nor the gene–environment independence assumption is necessary to derive the simple pseudolikelihood ℓ^*.

An alternative representation of l^* is very revealing. Consider a sampling scenario where each subject from the underlying population is selected into the case-control study using a Bernoulli sampling scheme where the selection probability for a subject given his or her disease status $D = d$ is proportional to $\mu_d = N_d/\text{pr}(D = d)$. Let $R = 1$ denote the indicator of whether a subject is selected in the case-control sample under the foregoing Bernoulli sampling scheme. With some algebra, one can

now show that the pseudolikelihood l^* can be expressed in the form

$$l^* = \sum_{i=1}^{N} \log\Bigg\{ \sum_{h^d \in \mathcal{H}_{G_i}} \mathrm{pr}(D_i | H_i^d = h^d, X_i, R_i = 1) \times \mathrm{pr}(H_i^d = h^d | X_i, R_i = 1) \Bigg\}$$
$$= \sum_{i=1}^{N} \log\{\mathrm{pr}(D_i, G_i | X_i, R_i = 1)\}. \quad (5)$$

When no environmental factors are involved, Stram et al. (2003) proposed an analysis of haplotype-based case-control studies using an "ascertainment-corrected joint likelihood" of the form $\prod_i \mathrm{pr}(D_i, G_i | R_i = 1)$. The representation of the l^* given in (5) suggests that under model (2) or (3) with $F(x)$ treated completely nonparametrically, the efficient retrospective maximum likelihood estimate of the haplotype frequency and the regression parameters can be obtained by conditioning on X in the approach of Stram et al. (2003).

In most parts of their article, Lin and Zeng considered retrospective maximum likelihood estimation under the model that assumes H^d and X are independent given G and then allows the distribution of $[X|G]$ to be completely nonparametric. This model has advantages and disadvantages. It is more flexible than the model (2) that assumes H^d and X are independent unconditionally; however, unlike model (3), it cannot allow direct association between haplotypes and environmental/demographic factors. Computationally, retrospective maximum likelihood assuming model (2) or model (3) completely avoids estimation of the distribution of the possibly high-dimensional covariates X. In contrast, under the model considered by Lin and Zeng, one must estimate the nonparametric distribution of X for each different genotype G, possibly stratified by subpopulations—a potentially daunting task. Finally, in situations where the gene–environment independence assumption is likely to be valid, either in the entire population or within subpopulations, based on results of Chatterjee and Carroll (2005) and Spinka et al. (2005), we conjecture that the retrospective maximum likelihood method assuming model (2) or model (3) can be much more efficient than that assuming the model of Lin and Zeng.

1.2 Prospective Methods for Retrospective Data

Lake et al. (2003) described methods for haplotype-based regression analysis based on the prospective likelihood of the data (D, G, X), ignoring the true case-control sampling design. For fixed values of the haplotype-frequency parameter θ, the score equations for the regression parameters $\beta^* = (\kappa, \beta_1)$ under model (2) corresponding to the prospective likelihood of the data is given by

$$0 = \sum_{i=1}^{N} \sum_{h^d \in \mathcal{H}_{G_i}} \frac{\partial}{\partial \beta^*} \log\{\mathrm{pr}_{\beta^*}(D_i | h^d, X_i)\} \times \mathrm{pr}_{\beta^*}(D_i | h^d, X_i) q(h^d; \theta) \times \Bigg(\sum_{h^d \in \mathcal{H}_{G_i}} \mathrm{pr}_{\beta^*}(D_i | h^d, X_i) q(h^d; \theta) \Bigg)^{-1}. \quad (6)$$

Unfortunately, this purely prospective score equation is biased under the case-control sampling design, even if the true haplotype frequencies were known and the underlying HWE and gene–environment independence assumptions were valid. However, a simple modification of the prospective score equation is unbiased,

$$0 = \sum_{i=1}^{N} \sum_{h^d \in \mathcal{H}_{G_i}} \frac{\partial}{\partial \beta^*} \log\{\mathrm{pr}_{\beta^*}(D_i | h^d, X_i)\} \times \mathrm{pr}_{\beta^*}(D_i | h^d, X_i) r_\Omega(h^d, X_i) q(h^d; \theta) \times \Bigg(\sum_{h^d \in \mathcal{H}_{G_i}} \mathrm{pr}_{\beta^*}(D_i | h^d, X_i) r_\Omega(h^d, X_i) q(h^d; \theta) \Bigg)^{-1}, \quad (7)$$

where

$$r_\Omega(h^d, X) = \frac{1 + \exp\{\kappa + m(h^d, x, \beta_1)\}}{1 + \exp\{\beta_0 + m(h^d, x, \beta_1)\}}.$$

Spinka et al. showed that with an appropriate rare disease approximation, the modified prospective estimating equation (7) is equivalent to the approximate estimating equation approach proposed by Zhao et al. (2003). Spinka et al. described strategies for estimating β_1 and κ based on the modified prospective estimating equation (7), where the nuisance parameters θ, and possibly β_0, are estimated based on score equation derived from the pseudolikelihood l^*. Simulation studies show that such a prospective approach generally tends to be much more robust to violation of both HWE and the gene–environment independence assumption compared with the retrospective maximum likelihood method (see also Satten and Epstein 2004).

2. COHORT–BASED STUDIES AND THE COX PROPORTIONAL HAZARDS MODEL

Lin and Zeng admirably describe fully efficient nonparametric maximum likelihood estimation for fitting a general haplotype-based semiparametric linear transformation model to cohort studies with unphased genotype data. An alternative estimator considered by Chen, Peters, Foster, and Chatterjee (2004) and Chen and Chatterjee (2005) for the popular Cox proportional hazard (CPH) model deserves attention. Consider the CPH model for specifying the hazard function for a subject given his or her diplotype status (H^d) and environmental covariates (X) as

$$\lambda[t | H^d, X] = \lambda_0(t) R(H^d, X; \beta_1), \quad (8)$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, $R(H^d, X; \beta_1)$ is a parametric function describing the relative risk associated with the exposure (H^d, X), and β_1 is the vector of associated regression parameters of interest. As before, assume that $\Pr(H^d) = q(H^d; \theta)$ is specified according to HWE and that θ denotes the associated haplotype frequency parameters. The model (8) cannot be used directly, because H^d is not observable. Following Prentice (1982), one can derive the hazard function for disease conditional on the observable genotype data G and

covariates X in the form

$$\lambda[t|G,X] = \lambda_0(t)R^*\{G,X;t,\beta_1,\theta,\Lambda_0(\cdot)\}, \quad (9)$$

where

$$\begin{aligned} &R^*\{G,X;t,\beta_1,\theta,\Lambda_0(\cdot)\} \\ &\quad = E\{R(H^d,X;\beta_1)|G,X,T\geq t\} \\ &\quad = \frac{\sum_{H^d\in\mathcal{H}_G} R(H^d,X;\beta_1)\,\mathrm{pr}[T>t|H^d,X]q(H^d;\theta)}{\sum_{H^d\in\mathcal{H}_G}\mathrm{pr}[T>t|H^d,X]q(H^d;\theta)}. \end{aligned}$$

In general, standard partial likelihood inference cannot be performed based on (9), because the relative risk function $R^*\{G,X;t,\beta_1,\theta,\Lambda_0(\cdot)\}$ itself depends on the baseline hazard function $\lambda_0(t)$. However, Chen et al. (2004) showed that an omnibus score test for genetic association can be performed using outputs from standard statistical software for partial likelihood analysis. Based on (9), Chen and Chatterjee (2005) also described alternative strategies for estimation of the risk parameters β_1. In particular, the authors observed that for rare disease, one could assume that $\mathrm{pr}[T>t|H^d,X]\approx 1$. The corresponding induced relative risk function,

$$R^*(G,X;\beta_1;\theta) = \frac{\sum_{H^d\in\mathcal{H}_G} R(H^d,X;\beta_1)q(H^d;\theta)}{\sum_{H^d\in\mathcal{H}_G} q(H^d;\theta)},$$

is free of the baseline hazard function $\lambda_0(t)$. Thus, under the rare disease approximation, one could estimate β by maximizing the partial likelihood associated with the relative risk function $R^*(G,X;\beta_1;\widehat{\theta})$, where $\widehat{\theta}$ is a consistent estimate of the haplotype-frequency parameters θ. Chen and Chatterjee described alternative strategies for obtaining consistent estimate of θ for cohort and nested case-control studies. A simple asymptotic variance estimator was also provided. Simulation studies for the full cohort design show that the loss of efficiency in this pseudolikelihood method was quite small compared with the fully efficient nonparametric maximum likelihood estimator (NPMLE) estimator proposed by Lin (2004).

An advantage of pseudolikelihood approach of Chen and Chatterjee (2005) is its wide applicability to alternative cohort-based study designs. In particular, for studies of rare diseases such as cancer, it is common to conduct case-control or case-cohort sampling within a cohort to select a subset of people for whom genotype and expensive environmental exposure information will be collected. Various alternative types of partial likelihoods that are currently available for analysis of nested case-control and case-cohort studies can be applied to estimate β_1 based on the induced relative risk function $R^*(G,X;\beta_1;\widehat{\theta})$. Future research is merited to study whether and how one can obtain the NPMLE for these alternative designs, especially when *both* genotype and environmental exposure data are available only for the selected subsample of the subjects. We look forward to Lin and Zeng's further innovations in this area.

ADDITIONAL REFERENCES

Chatterjee, N., and Carroll, R. J. (2005), "Semiparametric Maximum Likelihood Estimation in Case-Control Studies of Gene–Environment Interactions," *Biometrika*, 92, 399–418.

Chatterjee, N., Kalaylioglu, Z., and Carroll, R. J. (2005), "Exploiting Gene–Environment Independence in Family-Based Case-Control Studies: Increased Power for Detecting Associations, Interactions and Joint Effects," *Genetic Epidemiology*, 28, 138–156.

Chen, J., and Chatterjee, N. (2005), "Haplotype-Based Association Analysis in Cohort and Nested Case-Control Studies," *Biometrics*, in press.

Chen, J., Peters, U., Foster, C., and Chatterjee, N. (2004), "A Haplotype-Based Test of Association Using Data From Cohort and Nested Case-Control Epidemiologic Studies," *Human Heredity*, 58, 18–29.

Prentice, R. L. (1982), "Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model," *Biometrika*, 69, 331–342.

Spinka, C., Carroll, R. J., and Chatterjee, N. (2005), "Analysis of Case-Control Studies of Genetic and Environmental Factors With Missing Genetic Information and Haplotype-Phase Ambiguity," *Genetic Epidemiology*, 29, 105–127.

Comment

Jung-Ying TZENG and Kathryn ROEDER

All data analysis relies on a model that is, strictly speaking, not correct. Choices about which features to model and which to ignore distinguish successful models from the rest. Without artful modeling, statisticians would be unable to make inferences based on finite samples. In this wide-ranging article, Lin and Zeng (LZ hereinafter) make novel contributions to the statistical genetics literature by introducing new models and providing a rigorous statistical analysis of these models. Specifically, their article builds on a series of related works modeling the effect of haplotypes on the risk of disease. We congratulate the authors for providing a firm theoretical foundation in this exciting area of research. The authors investigate a family of models that address a broad range of sampling designs commonly used in genetic epidemiology, but for brevity we focus our remarks on those models appropriate to case-control data.

Schaid et al. (2002) published a practical methodological approach for haplotype association analysis using a prospective model to link the risk of disease to observed genetic data. The chosen model ignored two features of the data: the case-control sampling scheme that typically generates the data and poten-

Jung-Ying Tzeng is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: *jytzeng@stat.ncsu.edu*). Kathryn Roeder is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *roeder@stat.emu.edu*).

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000844